第 37 卷第 6 期 计算机应用研究 Vol. 37 No. 6
录用定稿 Application Research of Computers Accepted Paper

基于 HNC 句类的社区问答系统问句检索模型构建

王 宇，王 芳†

(大连理工大学 管理与经济学部，辽宁 大连 116024)

摘 要：社区问答系统中充斥着大量的噪声，给用户检索信息造成麻烦，以往的问句检索模型大多集中在词语层面。针对以上问题构建句子层面的问句检索模型。新模型基于概念层次网络(HNC)理论当中的句类知识，从句子的语用、语法和语义三个层面计算问句间相似度。通过问句分类算法确定查询问句和候选问句的问句类别，得到问句间的语用相似度，利用句类表达式的结构和语义块组成分别计算问句间的语法及语义相似度。在真实数据集上的实验表明基于 HNC 句类的新模型提高了问句检索结果的准确性。

关键词：社区问答系统；问句检索；HNC 理论；句类分析；相似度计算

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2018.11.0871

Construction of question retrieval model in community question answering system based on HNC sentence-category

Wang Yu, Wang Fang†

(*Faculty of Management & Economics*, *Dalian University of Technology*, *Dalian Liaoning* 116024, *China*)

Abstract: Community question answering system causes trouble for users to retrieve information due to useless information. Most of the previous question retrieval models focused on the word level. In order to solve the above problems, this paper proposes a question retrieval model at the sentence level. Based on the sentence-category of Hierarchical Network of Concept (HNC) theory, the new model calculated similarities between questions from the pragmatic, grammatical and semantic levels of the sentence. The model used the question classification algorithm to determine the categories of query question and candidate question, and thus obtained pragmatic similarity between questions. It used the sentence expression structure and the sentence semantic block to calculate grammatical and semantic similarities. Experiments on real data sets show that the new model based on HNC sentence-category improves the accuracy of question retrieval results.

Key words: community question answering system; question retrieval; hierarchical network of concept (HNC) theory; sentence category analysis; similarity calculation

0 引言

随着互联网的不断发展，网络上积累了大量的信息，普通用户可以通过搜索引擎获取想要的信息，但由于搜索引擎返回的是一系列相关文档而不是用户关心问题的答案，需要用户继续甄别信息，同时用户也不能用自然语言描述自己的问题，不能形成完整语义的稀疏关键词序列也导致搜索引擎的检索效果一般。

问答系统提供了一种新的用户问题到确切答案的信息检索过程，简化了不确定文档阅读查找的过程，一定程度上解决了搜索引擎的问题。随着 Web 2.0 的发展，普通用户逐渐从网络内容的接收者，变为网络内容的提供者，此时网络上充斥着大量的用户生成内容（user generated content, UGC）内容，而问答系统的主要参与者也变成了社区用户，形成了社区问答系统。以中文问答社区“知乎”为例，从开放至今已累计超过 1 000 万个提问以及 3 400 万个回答，大量的问题被解答，成为知识较为集中且有价值的网络资源。如何从这些大量的历史问答资源中找到与用户问题相似或者相关的内容，成为问答系统研究的一项主要内容。通过相似问题的检索可以减少用户获得答案的等待时间以及减轻相似内容重复提问造成的系统冗余的问题[1]。

社区问答系统的检索对象是问句和答案，它长度较一般文档短，存在数据稀疏的问题，并且内容还有自然语言表述随意，大部分词语有着一词多义和多词同义现象，致使无法通过词项严格匹配（如向量空间模型[2]）加以识别。

为了解决上述问题，Song 等人[3]将问题的语义信息和统计信息相结合，综合计算问句相似度大小；Cai 等人[4]使用潜在语义信息来解决问句检索中词汇空缺的问题，通过潜在语义信息消除词汇语义鸿沟；Jeon 等人[5]将语言模型应用到社区问答系统问句检索上，采用一元模型对社区型问答中的问答对进行建模，用于相似问句的发现工作；Xue 等人[6]在语言模型基础上提出了基于翻译模型的语言模型，较好地解决了检索过程中的词不匹配问题；文献[7]通过获得问句的潜在主题信息，提高翻译模型的检索性能。但由于翻译模型的准确性易受训练语料集及文本语义表达的影响，致使检索效果不够理想。夏远远等人[8]引入概念层次网络理论（hierarchical network of concept, HNC）中的词语知识库修正翻译概率，构建了新的问句检索模型，并给出问句检索模型的实现算法。

问句检索模型的着重点是考察句子的相似度[9~11]。之前的许多问句检索模型集中在问句的词汇匹配层面，如果忽视问句的语用特点，简单的将问句看做一系列词汇序列的集合，可能会导致根据问句检索得到的问题相应的答案并不是用户

收稿日期：2018-11-24；**修回日期：**2019-01-16

作者简介：王宇（1959-），男，吉林通化人，教授，硕导，博士，主要研究方向为文本挖掘、自然语言理解；王芳（1995-），女（满）（通信作者），辽宁葫芦岛人，硕士研究生，主要研究方向为文本挖掘、自然语言理解（705843314@qq.com）.

想要的。社区问答系统中具有用户社交属性以及问句的答案等丰富的可利用信息。文献[12]利用提问者的社交网络属性及查询问句内容构建问句的潜在表示，将问题内容文本信息与用户社交网络信息进行整合，对问题之间的相似性进行排序。问句类别是从一般语用的角度描述问句类型，属于同一个问题类别的问题比分属不同问题类别问题的更有可能具有相同的答案。文献[13]从词法、语法、语义三个层面提取问题的特征，再基于问题语法树[14]构建语义核函数，利用支持向量机(SVM)进行问题分类，提高了准确率。田卫东等人[15]认为问句相似度计算需在考虑问句本身相似度的基础上，同时需要考虑问句答案的相似度，而问句类别一定程度上可以限制问句的答案，因此在问句相似度计算时加入了问句类型相似度计算这个维度，提供了问句相似度计算的新思路。

本文采用 HNC 理论的句类分析方法获取问句更深层次的语法语义信息，从计算机自动化角度出发，利用层次分析的方法确定问句类型，在句子层面分别从语法、语义和语用三个方面共同构建问句检索模型，提高检索模型的准确度。

1　HNC 句类分析

HNC(hierarchical network of concept)理论是中国科学院声学所黄曾阳研究员提出的一种自然语言处理体系[16]，它面向整个自然语言来描述大脑认知结构的具体模式，将认知结构分为局部和全局两个联想脉络，其中局部联想脉络是词汇层面的概念表述体系，全局联想脉络是语句及篇章层面的联想[17]，主要包括语义块和句类理论，语义块从语言深层描述一个句子，解决了从词或者短语层面难以界定句子语义的问题。任意自然语言句子都可以用一种形式上的句类表达式代表，句类表达式由语义块组成，即语义块是句类的函数。常见的 HNC 句类有作用句、过程句、转移句等。句类表达式从语法层面揭示了句子的语义块序列，相似度高的句子往往具有相同的语义块序列。因此问句语法层面的相似度大小可以由问句句类表达式相似度得出。语义块可以是一个词汇或者词汇序列，甚至是句子，但都是由词汇组成，语义块的相似度度量就可以通过词汇语义相似度代表，HNC 理论在词汇层面也有一套词汇概念语义网络，在表达词汇语义的完备性上有比较优异的表现，问句语义层面相似度可以由词汇的概念语义来确定。

HNC 构建了一套新式的自然语言处理体系，HNC 系统在语义分析中可以通过句类分析完成完整的语义分析过程，将获得包括主辅语义块在内的一系列句类知识，这样的句类知识为计算机自动化的句子相关计算提供了便利的条件。句类分析得到的句类表达式是句法层面的句子分析，HNC 聚类共有 57 种基本句类以及 3 192 种混合句类，完整地描述几乎全部自然语言的句子语法搭配现象。HNC 组成句类表达式的相应语义块则包含相应的语义信息。HNC 将句子下一级语义构成单位命名为语义块。语义块的范围很广，可以是一个词或者一个短语甚至于一个句子。语义块按照在句子中起的作用分为主语义块和辅语义块。主语义块分别命名为特征要素、作用者、对象和内容；辅语义块有七种，分别为手段、工具、途径、比照、条件、原因、结果和目的。在利用 HNC 理论做句类分析的工作中，陈鸿[18]将 HNC 句类分析分为语义块感知、句类假设、句类检验以及语义块构成分析等多个模块，并实现了完整的 HNC 句类分析的算法。池哲洁[19]将句类分析的结果分为主辅语义块及表层表达式相似度层面和深层语义层面计算句子相似度，但并未考虑辅语义块对句子相似度的贡献。史燕[20]通过句类分析，将语义块对应计算相似度，最后按照平均的方式计算最终句子的相似度。该方法虽然利用了 HNC 理论的相关知识，但忽略了语义块构成上的多种因素，因此是不全面的。

HNC 句类分析的最终目的是为了获得句子的句类表达式，但由于 HNC 句类分析的有效性需要依赖于 HNC 知识库，文献[18]所提出的句类分析算法在实际操作中存在一定的局限性。考虑到句类是由语义块组成，语义块作为语义的底层构成单位，在承担语义作用的基础上，也有其语用层面上的作用，利用已知句子 HNC 句类表达式，通过词性标注序列的相似性推导另一未知句子的句类表达式。对于句子“美国军方向波斯尼亚战争的受害者空投救援物资”，作为一种物转移句有句类表达式 $T2J = T2A + l + REC + T2 + T2C$，其中 $T2A$ 表示物转移的发起者(美国　军方)，l 是转移的方向，REC 代表了转移接收者(波斯尼亚　受害者)，$T2$ 为特征要素(空投)，$T2C$ 则表示转移内容(救援物资)。同时该句有词性标注序列如：ns(美国)，n(军方)，p(向)，ns(波斯尼亚)，n(战争)，n(受害者)，v(空投)，n(救援物资)，有如表 1 所示的对应关系，这样建立了句类表达式和句子的词性序列的关系。

通过收集已经进行 HNC 句类分析的句子，获取句类表达式，之后进行句子词汇序列的词性标注，将这部分句子作为先验知识，通过句子词性序列的相似度来获得没有进行句类分析句子的句类表达式。

接下来将利用 HNC 句类分类分析获得问句语法和语义两部分句子信息以及前面提到的问句类型代表的语用信息，综合构建问句检索模型。

表 1　语义块与词性序列的对应

Table 1　Correspondence between semantic block and part of speech sequence

语义块	词性序列
$T2J$	ns , n
l	p
REC	ns , n
$T2$	v
$T2C$	n

2　基于 HNC 句类分析的问句检索模型

根据前文的问句类别和问句语义的相关分析，从语用、语法和语义三个层次综合构建问句检索模型。检索模型的有效性需要通过检索结果排序的准确性来验证，一般在信息检索的任务中，通过 MRR、MAP、AP@1 等指标来说明检索效果的好坏。

AP@1(average precision)：关于特定查询的检索排序结果中，相关问句在第一位的平均百分比。

MAP(mean average precision)：表示返回结果的平均准确率，本实验中计算每个查询返回的前 10 个结果的平均准确率，即 MAP10。

MRR(mean reciprocal rank)：在保证 MAP 检索准确性的同时，还要关注检索结果的排序顺序，MRR 是加入排序顺序影响后的检索结果准确率。

设计一个代表检索模型的排序公式，是验证模型有效性的一个必不可少的步骤。本章设计了一种检索模型的排序公式，新的问句检索的排序机制如下：

$$Sim_{mss}(Q,C) = \alpha Sim_{cat}(Q,C) + \beta Sim_{gse}(Q,C) + \gamma Sim_{sem}(Q,C) \quad (1)$$

其中：Q,C 分别代表了问句检索中的查询问句和候选问句；Sim_{mss} 表示 Q,C 的总体相似度；Sim_{cat}、Sim_{gse}、Sim_{sem} 分别代表 Q,C 的问句类型相似度、语法相似度以及语义相似度；α、β、

γ 分别为各分项的调节参数，表示问句查询中三种相似度所占权重。算法流程如图 1 所示。

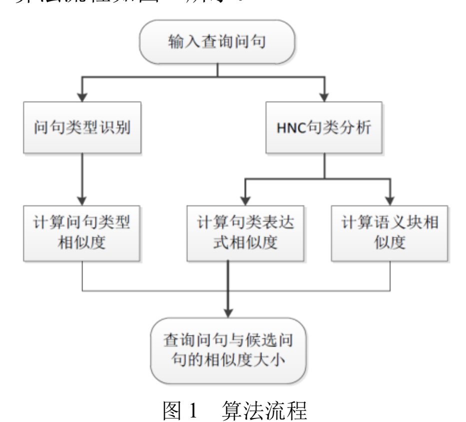


图 1　算法流程

Fig. 1　Algorithm flowchart

下面分别介绍各部分的相似度计算过程。

2.1　问题类型相似度度量

传统问答系统的中的问题多是以事实类问题呈现，如“第一位获得诺贝尔奖的中国人是谁”“什么是防火墙”等，但社区问答系统中的问题从语用的角度看有多种问题形式，文献[21]将社区问答系统的问题形式分为了定义、事实、过程、原因、观点、是非和描述等七种类型。问句的类型确定了相关问题的答案焦点，如过程类问题答案和定义类问题答案是完全不同的，由于问句检索的目的是在历史问答中找到与用户新提交问题相似的问题，进而获得新提交问题的答案，因此问题类型所代表语用信息对问句检索也有着比较重要的意义。

关于中文文本分类的研究中，一般都是通过提取分类特征进行分类。张振豪等人[22]在文献[23]提出将关键词应用到文档分类的基础上，在短文本分类当中加入关键词相似度计算，选择 K-近邻(KNN)和 SVM 作为分类器，较传统分类方法提升了分类效果。高超等人[24]在词袋特征基础上提出融合中心词、主语、疑问词以及疑问词相关成分等问题类别线索词集，从问句的句法或语义信息角度挖掘更多对分类有用的信息。余本功等人[25]提出问句中的疑问词是具有重要影响的特征，可以通过构造疑问词注意力矩阵强化模型对疑问词特征的重点关注，从而提高问句分类的准确率。文献[26]提取问题的主干词和疑问词作为分类特征，之后使用贝叶斯分类器对问句进行分类。本节在文献[26]的特征维度之上，考虑社区问答系统的人工标记属性，将问题的分类标签作为问句分类的另一特征，采用贝叶斯分类器对问句进行分类。

首先确定待检索问句和候选问句分别属于的问题类型。问句分类方法按照上述的提取分类特征，输入贝叶斯分类器进行分类。之后在计算问句类型相似度时，考虑到问句检索的目的是获取问答系统中已解决问题的答案，分属不同问题类型的答案虽然焦点可能不一样，但答案的背景知识或许对提问者有一定的参考价值，因此问题类型相似度由式(2)确定如下：

$$Sim_{cat}(Q,C)=\begin{cases}1 & Q,C\text{ 同属一个小类问题类型}\\ \alpha_1 & Q,C\text{ 同属一个大类问题类型}\\ 0 & Q,C\text{ 属于不同的问题类型}\end{cases} \tag{2}$$

其中：α_1 是不同问题类型的相似系数,在参数调优阶段确定。

综上问句类别相似度计算步骤如下:

算法 1　问句类别相似度

输入：问句 Q,C 。

输出：问句类别相似度 $Sim_{cat}(Q,C)$ 。

a）分别提取问句 Q,C 的焦点词集合；

b）将 Q,C 的焦点词集合与表示问题类别的词语集合做语义相似度比较，判断问句 Q,C 所属的问句类别；

c）采用式(2)计算问句类别相似度 $Sim_{cat}(Q,C)$ 。

2.2　问句句类表达式相似度度量

HNC 句类分析的结果可以得到句子下层的组成要素，即语义块。在考察由语义块组成的句类表达式上的相似度时，比较待检索问句与候选问句之间的语义块组成，采用量化的方法计算句类表达式的相似度。语义块是句类的函数，在句子中起核心部分的语义块称为主语义块，说明部分的语义块称为辅语义块。相似度程度高的句子往往具有较多的相同语义块，定义如下句类表达式相似度计算的公式：

$$Sim_{gse}(Q,C)=\beta_1 Sim_{gsem}(Q,C)+\beta_2 Sim_{gsef}(Q,C) \tag{3}$$

其中：β_1,β_2 为调节参数，表示主辅语义块相似程度在问句表达式相似度计算中所占权重；Sim_{gsem} 、Sim_{gsef} 分别代表主辅语义块相似度。借鉴文献[18]中的计算方法，按照主辅语义块的类别确定相似度大小，其中：

$$Sim_{gsem}(Q,C)=\begin{cases}1 & Q,C\text{ 具有相同的句类符号}\\ \beta_{11} & Q,C\text{ 同属于广义作用或效用句}\\ 0 & \text{其他情况}\end{cases} \tag{4}$$

$$Sim_{gsef}(Q,C)=\begin{cases}1 & FK_1=FK_2\ \ AND\ \ C(FK_1)=C(FK_2)\\ \beta_{21} & FK_1=FK_2\ \ AND\ \ C(FK_1)\neq C(FK_2)\\ \beta_{22} & FK_1\cap FK_2\neq\varnothing\ \ AND\ \ C(FK_1)\neq C(FK_2)\\ 0 & FK_1\cap FK_2\neq\varnothing\end{cases} \tag{5}$$

其中：系数 $\beta_{11},\beta_{21},\beta_{22}$ 表示不同情形下主辅语义块对应的相似程度，均在参数调优阶段确定；FK_i 表示第 i 个句子的辅语义块集合；$C(FK_i)$ 表示第 i 个句子辅语义块的内容集合。综合以上分析，问句表达式相似度计算步骤如下：

算法 2　问句句类表达式相似度

输入：问句 Q,C 。

输出：问句句类表达式相似度 $Sim_{gse}(Q,C)$ 。

a）对 Q,C 做 HNC 句类分析，得到相应的句类表达式集合；

b）分别采用式(4)和式(5)计算主辅语义块相似度 $Sim_{gsem}(Q,C)$ ，$Sim_{gsef}(Q,C)$ ；

c）将 $Sim_{gsem}(Q,C)$ ，$Sim_{gsef}(Q,C)$ 代入式(3)得到最终的 Q,C 问句句类表达式相似度。

2.3　问句语义块相似度度量

2.2 节的问句句类表达式度量相关计算方法，是句子表层的相似度计算方法，本节根据 HNC 句类分析的结果获得语义块计算句子语义层面的相似度。

语义块可以是词语、短语或者是句子，但都是词汇序列，可以利用文献[8]中介绍的方法计算词汇序列的相似度。句类分析得到的语义块有主语义块和辅语义块两种。其中主语义块对句子语义起到支配作用，辅语义块往往表示说明部分。一般关于句子相似度的计算忽略辅语义块对句子的贡献，这在一般自然语言句子层面是合理的，但在社区问答系统中，由于问句类型有限，大量的问题具有相同的句类表达式，这时辅语义块对于衡量句子相似度就至关重要。

在计算问句语义相似度时将问句语义块一一对应起来，分别求对应语义块的相似度大小。计算公式如下：

$$Sim_{sem}(Q,C)=\gamma_1 Sim_{semm}(Q,C)+(1-\gamma_1)Sim_{semf}(Q,C) \tag{6}$$

其中：$Sim_{semm}(Q,C)$ 表示 Q,C 对应的主语义块的相似度大小；$Sim_{semf}(Q,C)$ 表示 Q,C 对应的辅语义块的相似度大小。因为主语义块对句子语义起主要作用，系数 γ_1 表示主语义块相似度所占权重，应设置为大于 0.5 的值。综上所述问句语义块相

似度计算步骤如下：

算法 3 问句语义块相似度

输入：问句 Q,C 。

输出：问句 Q,C 的语义块相似度 $Sim_{sem}(Q,C)$ 。

a）将句类分析得到的 Q,C 主辅语义块对应起来；

b）采用文献[8]中提出词汇语义相似度计算方法，分别得到 Q,C 主辅语义块之间的相似度 $Sim_{semm}(Q,C)$ ， $Sim_{semf}(Q,C)$ ；

c）最后将 $Sim_{semm}(Q,C)$ ， $Sim_{semf}(Q,C)$ 代入式(6)计算最终的 Q,C 的语义块相似度 $Sim_{sem}(Q,C)$ 。

至此问句检索模型的排序公式各部分已经分别计算完毕，将得到 $Sim_{cat}(Q,C)$ 、 $Sim_{gse}(Q,C)$ 以及 $Sim_{sem}(Q,C)$ 代入式(1)组成完整的问句检索模型排序公式。下面通过实验验证本章提出算法的有效性。

3 实验

为了证明第 2 章中提出的问句检索模型的有效性，本文选取了较为经典的三个检索模型，即基于向量空间模型的检索模型、基于语言模型的检索模型以及基于翻译模型的检索模型做比较分析。其中基于语言模型的检索模型的算法是由文献[5]具体实现的，基于翻译模型的检索模型的算法由文献[6]具体实现的。选取以上三个模型是因为这三个模型简单有效，以往大量的研究选择以上三个模型作为模型基础，进而设计检索模型。选择的语言模型以及翻译模型的具体实现文献[5]和文献[6]，是因为它们代表两种检索模型公认的具有较高检索性能的实现，将其作为参考模型增强了实验结果的可靠性。

在 HNC 理论概念词库建设的限制下，本实验的数据集语言环境为中文。首先需要为实验中采用的对比模型翻译模型构建训练语料集，以获取先验知识。翻译模型的训练语料集构建过程：首先从知乎问答社区上随机收集问答对。之后通过问答系统对问答资源的标记特征找出这些问答对中的相似问句，利用相似问句训练翻译。测试数据集部分，使用已经被人工标记的 1 140 个测试问句，每个测试问句有 20 个候选问句，候选问句分别被标记为相似或者不相似；同时选择 20 个测试问句作为模型参数调试集，通过调整得到第 2 章中的各参数设置，如表 2 所示。最后利用剩下的测试问句验证各对比模型。

表 2 模型参数设置

Table 2 Model parameter settings

α	β	γ	α_1	β_1	β_2	β_{11}	β_{21}	β_{22}	γ_1
0.3	0.3	0.4	0.5	0.7	0.3	0.5	0.7	0.4	0.8

表 3 表示的是各模型在各评价指标上的对比。其中 Improvement（IMPR）表示各模型在各指标上相较于向量空间模型提升值。

从表 3 的实验结果可以看出，简单的向量空间模型在各检索指标上均落后于其他模型，说明了基于词项严格词匹配的一类模型由于面临数据稀疏的问题，导致大部分相似问句不能够被召回。这样以向量空间模型为代表的词袋模型在用户生成内容丰富的文本资源环境下，检索有效性不如大规模文档资源。

此外本文也可以发现，翻译模型和语言模型在检索指标上提升度比较明显，是因为翻译模型或者语言模型均具有较完善的平滑机制，其检索效果也相应优于普通的向量空间模型。同时，翻译模型的检索有效性也优于语言模型，是因为一方面翻译模型可以通过规模语料集获得语言模型不具备的背景语义知识，另外翻译模型通过翻译概率代替一般文本相似度计算的词汇相似度部分，解决了词汇鸿沟问题给相似度计算带来的障碍，提升了检索效果。

表 3 各模型在 MRR、MAP、AP@1 上的对比

Table 3 Comparison of models on MRR, MAP, and AP@1

模型	AP@1	MAP	MRR
向量空间模型	0.21	0.32	0.35
IMPR	N/A	N/A	N/A
语言模型	0.52	0.56	0.67
IMPR	1.4762	0.75	0.9143
翻译模型	0.61	0.60	0.73
IMPR	1.9048	0.88	1.0857
HNC 句类分析模型	0.67	0.72	0.77
IMPR	2.1905	1.25	1.2

最后可以看出本文提出的基于 HNC 句类分析的方法较以往的方法在问句检索效果上有更好的提升。可以发现另外三种模型都是将问句检索中的问句相似度进行简化计算，使得计算的问句相似度计算结果只能代表问句语义或者词义相似度，并不是全面的语义、语法及语用的完整相似度，因此使得检索结果失去一部分相似问句。实验表明本文提出的基于句法分析的检索模型，在结合语义、语法和语用的多角度的相似度后，提高了问句检索的效果，说明本文方法的有效性。在实际的实验验证过程中，基于 HNC 句类分析模型较前三种模型计算过程存在一定的复杂性，最终相似度计算维度较多，使得模型的可靠性更高，个别维度相似度的误差对最终结果的准确性影响较小。

4 结束语

借助 HNC 自然语言处理理论的句类分析方法，提出了一种新的基于句类分析的结果，从语法、语义和语用三个角度构建问句检索模型，较之前的模型提高了问句检索的性能。对于给定的一个查询，首先通过问句分类算法确定查询问句和候选问句的问句类别，进而获得问句类型相似度，即语用相似度；之后对查询问句和候选问句做 HNC 句类分析，将得到的句类分析结果从语法和语义两个方面加以利用，得到最终的问句相似度，将问句相似度作为问句检索的依据，对候选问句进行排序。实验表明了所提出的方法有效性显著高于之前的经典检索模型。本文在问句的语用层面考虑了问句类别，社区问答系统当中的用户社交属性也十分重要，下一步工作考虑是否可以将用户需求、兴趣考虑到语用相似度计算当中。

参考文献：

[1] 延霞，范士喜. 基于问答社区的海量问句检索关键技术研究 [J]. 计算机应用与软件, 2013, (7): 315-317. (Yan Xia, Fan Shixi. On key techniques of massive question sentences retieval based on community QA [J]. Computer Applications and Software, 2013, (7): 315-317.)

[2] 苏小虎，杨思春. 基于改进 VSM 的中文问答系统研究 [J]. 情报理论与实践, 2008, 31 (4): 624-627. (Su Xiaohu, Yang Sichun. Research on Chinese question answering system based on improved VSM [J]. Information Studies: Theory & Application, 2008, 31 (4): 624-627.)

[3] Song W, Feng M, Gu N, *et al.* Question similarity calculation for FAQ answering [C]// Proc of International Conference on Semantics, Knowledge and Grid. 2007: 298-301.

[4] Cai L, Zhou G, Liu K, *et al.* Learning the latent topics for question retrieval in community QA [C]// Proc of International Joint Conference on Natural Language Processing. 2015.

[5] Jeon J, Croft W B, Lee J H. Finding similar questions in large question and answer archives [C]// Proc of ACM International Conference on Information and Knowledge Management. 2005: 84-90.

[6] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008: 475-482.

[7] 张伟男，张宇，刘挺. 一种面向社区型问句检索的主题翻译模型 [J]. 计算机学报, 2015, 38 (2): 313-321. (Zhang Weinan, Zhang Yu, Liu Ting. A topic inference based translation model for question retrieval in community-based question answering services [J]. Chinese Journal of Computers, 2015, 38 (2): 313-321.)

[8] 夏远远，王宇. 基于 HNC 理论的社区问答系统问句检索模型构建 [J]. 计算机应用与软件, 2018, 35 (8): 98-101,169. (Xia Yuanyuan, Wang Yu. Construction of a query model of community question answering system based on HNC theory [J]. Computer Applications and Software, 2018, 35 (8): 98-101,169.)

[9] 熊大平，王健，林鸿飞. 一种基于 LDA 的社区问答问句相似度计算方法 [J]. 中文信息学报, 2012, 26 (5): 40-45. (Xiong Daping, Wang Jian, Lin Hongfei. An LDA-based approach to finding similar questions for community question answer [J]. Journal of Chinese Information Processing, 2012, 26 (5): 40-45.)

[10] 陈康，樊孝忠，刘杰，等. 基于问句语义表征的中文问句相似度计算方法 [J]. 北京理工大学学报, 2007, 27 (12): 1073-1076. (Chen Kang, Fan Xiaozhong, Liu Jie, *et al.* Calculation method of Chinese question semantic similarity based on question semantic representation [J]. Transactions of Beijing Institute of Technology, 27 (12): 1073-1076.)

[11] 杨海天. 社区问答系统中问句检索技术的研究 [D]. 大连：大连理工大学, 2014. (Yang Haitian. The study on question retrieval technology in community question answer system [D]. Dalian: Dalian University of Technology, 2014.)

[12] Chen Z, Zhang C, Zhao Z, *et al.* Question retrieval for community-based question answering via heterogeneous social influential network [J]. Neurocomputing, 2018, 285: 117-124.

[13] 江龙泉，张波，胡志鹏，等. 问答系统中基于语义核函数的问题分类算法 [J]. 上海师范大学学报：自然科学版, 2018 ，47(1): 53-56. (Jiang Longquan, Zhang Bo, Hu Zhipeng, *et al.* A semantic kernel function based question classification algorithm in question answering system [J]. Journal of Shanghai Normal University: Natural Sciences, 2018,47(1): 53-56..)

[14] Mishra A, Jain S K. A survey on question answering systems with classification [J]. Journal of King Saud University: Computer and Information Sciences, 2016, 28 (3): 345-361.

[15] 田卫东，强继朋. 基于问句类型的问句相似度计算 [J]. 计算机应用研究, 2014, 31 (4): 1090-1093. (Tian Weidong, Qiang Jipeng. Questions similarity computation based on question classification [J]. Application Research of Computers, 2014, 31 (4): 1090-1093.)

[16] 黄曾阳. HNC（概念层次网络）理论 [M]. 北京：清华大学出版社, 1999. (Huang Zengyang. HNC (hierarchical network of concept) theory [M]. Beijing: Tsinghua University Press, 1999.)

[17] 黄曾阳. HNC 理论概要 [J]. 中文信息学报, 1997, 11 (4): 11-20. (Huang Zengyang. HNC theory summary [J]. Journal of Chinese Information Processing, 1997, 11 (4): 11-20.)

[18] 陈鸿. 自然语言理解——基于 HNC 理论的句类分析研究 [D]. 长春：长春理工大学, 2004. (Chen Hong. Natural language understanding-the study of sentence category analysis based on the theory of HNC [D]. Changchun: Changchun University of Science and Technology, 2004.)

[19] 池哲洁. 利用语言概念知识的事件文本分析关键技术研究 [D]. 北京：中国科学院声学所, 2017. (Chi Zhejie. Research on key techniques of event text analysis using language concept knowledge [D]. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2017.)

[20] 史燕. 基于 HNC 的汉语句子相似度算法的研究 [D]. 镇江：江苏大学, 2009. (Shi Yang. The research on chinese sentence similarity algorithm based on HNC [D]. Zhenjiang: Jiangsu University, 2009.)

[21] 董才正，刘柏嵩. 面向问答社区的中文问题分类 [J]. 计算机应用, 2016, 36 (4): 1060-1065. (Dong Caizheng, Liu Baisong. Community question answering-oriented Chinese question classification [J]. Journal of Computer Applications, 2016, 36 (4): 1060-1065.)

[22] 张振豪，过弋，韩美琪，等. 基于关键词相似度的短文本分类方法研究 [J/OL]. 计算机应用研究：1-6 [2018-11-22]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 04. 0440. (Zhang Zhihao, Guo Yi, Han Meiqi, *et al.* Research on short text classification based on keyword similarity [J/OL]. Application Research of Computers: 1-6 [2018-11-22]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 04. 0440.)

[23] Onan A, Korukoğlu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification [J]. Expert Systems with Applications, 2016, 57 (9): 232-247.

[24] 高超，杨思春，万家山. 融合类别线索词的中文问题分类 [J]. 苏州科技学院学报：自然科学版, 2018 35 (2) : 73-78. (Gao Chao, Yang Sichun, Wan Jiashan. Chinese question classification based on fusion category clue words [J]. Journal of Suzhou University of Science and Technology: Natural Science, 2018 35 (2): 73-78.)

[25] 余本功，许庆堂，张培行. 基于 MAC-LSTM 的问题分类研究 [J/OL]. 计算机应用研究：1-6 [2018-11-22]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 05. 0452. (Yu Bengong, Xu Qingtang, Zhang Peihang. Question classification based on MAC-LSTM [J/OL]. Application Research of Computers: 1-6 [2018-11-22]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 05. 0452.)

[26] 文勖，张宇，刘挺，等. 基于句法结构分析的中文问题分类 [J]. 中文信息学报, 2006, 20 (2): 33-39. (Wen Xu, Zhang Yu, Liu Ting, *et al.* Syntactic structure parsing based chinese question classification [J]. Journal of Chinese Information Processing, 2006, 20 (2): 33-39.)